INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 :
G10L 9/00

A1

(11) International Publication Number: WO 94/20952

(43) International Publication Date: 15 September 1994 (15.09.94)

(21) International Application Number: PCT/US94/02542

(22) International Filing Date: 8 March 1994 (08.03.94)

(30) Priority Data:
08/032,850 12 March 1993 (12.03.93) US

(71) Applicant: SRI INTERNATIONAL [US/US]; 333 Ravenswood Avenue, Menlo Park, CA 94025 (US).

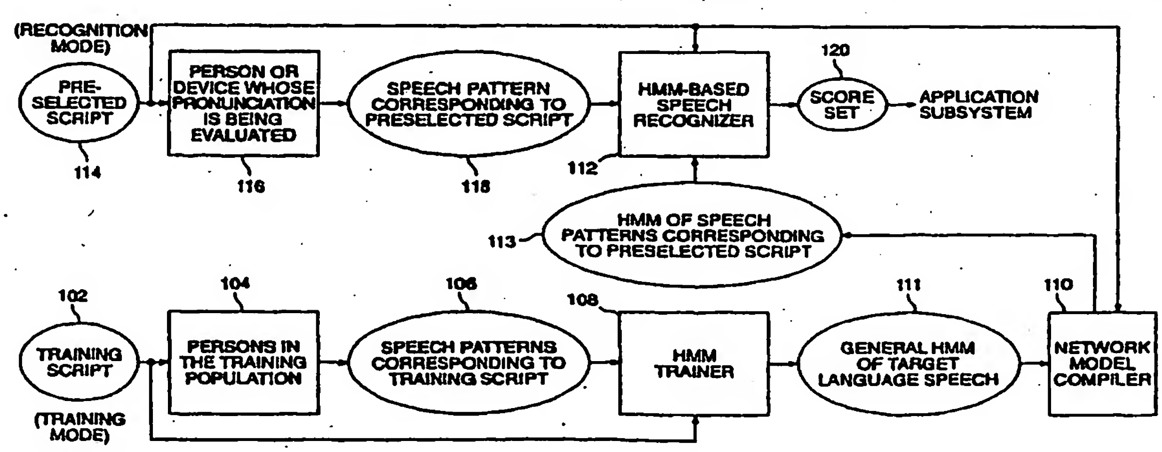
(72) Inventors: RTISCHEV, Dimitry; 675 Sharon Park Drive, #222, Menlo Park, CA 94025 (US). BERNSTEIN, Jared, C.; 1330 Tasso Street, Palo Alto, CA 94301 (US). CHEN, George, T.; 723 Marsh Road, #7, Menlo Park, CA 94025 (US). BUTZBERGER, John, W.; 736 Bounty Drive, #3609, Foster City, CA 94404 (US).

(74) Agent: ALLEN, Kenneth, R.; Townsend and Townsend Khourie and Crew, Steuart Street Tower, 20th floor, One Market Plaza, San Francisco, CA 94105 (US).

(81) Designated States: CA, JP, KR, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: METHOD AND APPARATUS FOR VOICE-INTERACTIVE LANGUAGE INSTRUCTION



(57) Abstract

Spoken-language instruction method and apparatus employ context-based speech recognition for instruction and evaluation. A finite state grammar set (113) corresponding to the range of word sequence patterns in the lesson is employed as a constraint on a hidden Markov model (HMM) search apparatus in an HMM speech recognizer (112). The invention includes a system with an interactive decision mechanism which employs at least three levels of error tolerance to simulate a natural level of patience in human-based interactive instruction. A linguistically-sensitive utterance endpoint detector is provided for judging termination of a spoken utterance to simulate human turn-taking in conversational speech.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgystan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo			SE	Sweden
CH	Switzerland	KR	Republic of Korea	SI	Slovenia
CI	Côte d'Ivoire	KZ	Kazakhstan	SK	Slovakia
CM	Cameroon	LI	Liechtenstein	SN	Senegal
CN	China	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
CZ	Czech Republic	LV	Latvia	TJ	Tajikistan
DE	Germany	MC	Monaco	TT	Trinidad and Tobago
DK	Denmark	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	US	United States of America
FI	Finland	ML	Mali	UZ	Uzbekistan
FR	France	MN	Mongolia	VN	Viet Nam
GA	Gabon				

METHOD AND APPARATUS FOR VOICE-INTERACTIVE LANGUAGE INSTRUCTION

BACKGROUND OF THE INVENTION

This invention relates to speech recognition and more particularly to the types of such systems based on a hidden Markov models (HMM) for use in language or speech instruction.

By way of background, an instructive tutorial on hidden Markov modeling processes is found in a 1986 paper by Rabiner et al., "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, Jan. 1986, pp. 4-16.

Various hidden-Markov-model-based speech recognition systems are known and need not be detailed herein. Such systems typically use realizations of phonemes which are statistical models of phonetic segments (including allophones or, more generically, phones) having parameters that are estimated from a set of training examples.

Models of words are made by making a network from appropriate phone models, a phone being an acoustic realization of a phoneme, a phoneme being the minimum unit of speech capable of use in distinguishing words. Recognition consists of finding the most-likely path through the set of word models for the input speech signal.

Known hidden Markov model speech recognition systems are based on a model of speech production as a Markov source. The speech units being modeled are represented by finite state machines. Probability distributions are associated with the transitions leaving each node, specifying the probability of taking each transition when visiting the node. A probability distribution over output symbols is associated with each node. The transition probability distributions implicitly model duration. The output symbol distributions are typically used to model speech signal characteristics such as spectra.

The probability distributions for transitions and output symbols are estimated using labeled examples of speech.

Recognition consists of determining the path through the Markov network that has the highest probability of generating the observed sequence. For continuous speech, this path will correspond to a sequence of word models.

Models are known for accounting for out-of-vocabulary speech, herein called reject phone models but sometimes called "filler" models. Such models are described in Rose et al., "A Hidden Markov Model Based Keyword Recognition System," Proceedings of IEEE ICASSP, 1990.

The specific hidden Markov model recognition system employed in conjunction with the present invention is the Decipher speech recognizer, which is available from SRI International of Menlo Park, California. The Decipher system incorporates probabilistic phonological information, a trainer capable of training phonetic models with different levels of context dependence, multiple pronunciations for words, and a recognizer. The co-inventors have published with others papers and reports on instructional development peripherally related to this invention. Each mentions early versions of question and answer techniques. See, for example, "Automatic Evaluation and Training in English Pronunciation," Proc. ICSLP 90, Nov. 1990, Kobe, Japan. "Toward Commercial Applications of Speaker-Independent Continuous Speech Recognition," Proceedings of Speech Tech 91, (April 23, 1991) New York, New York. "A Voice Interactive Language Instruction System," Proceedings of Eurospeech 91, Genoa, Italy September 25, 1991. These papers described only what an observer of a demonstration might experience.

Other language training technologies are known. For example, U.S. Pat. No. 4,969,194 to Ezawa et al. discloses a system for simple drilling of a user in pronunciation in a language. The system has no speech recognition capabilities, but it appears to have a signal-based feedback mechanism using a comparator which compares a few acoustic characteristics of speech and the fundamental frequency of the speech with a reference set.

U.S. Pat. No. 4,380,438 to Okamoto discloses digital controller of an analog tape recorder used for recording and

playing back a user's own speech. There are no recognition capabilities.

U.S. Patent No. 4,860,360 to Boggs is a system for evaluating speech in which distortion in a communication channel is analyzed. There is no alignment or recognition of the speech signal against any known vocabulary, as the disclosure relates only to signal analysis and distortion measure computation.

U.S. Patent No. 4,276,445 to Harbeson describes a speech analysis system which produces little more than an analog pitch display. It is not believed to be relevant to the subject invention.

U.S. Patent No. 4,641,343 to Holland et al. describes an analog system which extracts formant frequencies which are fed to a microprocessor for ultimate display to a user. The only feedback is a graphic presentation of a signature which is directly computable from the input signal. There is no element of speech recognition or of any other high-level processing.

U.S. Patent No. 4,783,803 to Baker et al. discloses a speech recognition apparatus and technique which includes means for determining where among frames to look for the start of speech. The disclosure contains a description of a low-level acoustically-based endpoint detector which processes only acoustic parameters, but it does not include higher level, context-sensitive end-point detection capability.

What is needed is a recognition and feedback system which can interact with a user in a linguistic context-sensitive manner to provide tracking of user-reading of a script in a quasi-conversational manner for instructing a user in properly-rendered, native-sounding speech.

SUMMARY OF THE INVENTION

According to the invention, an instruction system is provided which employs linguistic context-sensitive speech recognition for instruction and evaluation, particularly language instruction and language fluency evaluation. The system can administer a lesson, and particularly a language

lesson, and evaluate performance in a natural voice-interactive manner while tolerating strong foreign accents from a non-native user. The lesson material and instructions may be presented to the learner in a variety of ways, including, but not limited to, video, audio or printed visual text. As an example, in one language-instruction-specific application, an entire conversation and interaction may be carried out in a target language, i.e., the language of instruction, while certain instructions may be in a language familiar to the user.

In connection with preselected visual information, the system may present aural information to a trainee. The system prompts the trainee-user to read text aloud during a reading phase while monitoring selected parameters of speech based on comparison with a script stored in the system. The system then asks the user certain questions, presenting a list of possible responses. The user is then expected to respond by reciting the appropriate response in the target language. The system is able to recognize and respond accurately and in a natural manner to scripted speech, despite poor user pronunciation, pauses and other disfluencies.

In a specific embodiment, a finite state grammar set corresponding to the range of word sequence patterns in the lesson is employed as a constraint on a hidden Markov model (HMM) search apparatus in an HMM speech recognizer which includes a set of hidden Markov models of target-language narrations (scripts) produced by native speakers of the target language.

The invention is preferably based on use of a linguistic context-sensitive speech recognizer, such as the Decipher speech recognizer available from SRI International of Menlo Park, California, although other linguistic context-sensitive speech recognizers may be used as the underlying speech recognition engine.

The invention includes a mechanism for pacing a user through an exercise, such as a reading exercise, and a battery of multiple-choice questions using an interactive decision mechanism. The decision mechanism employs at least three

levels of error tolerance, thereby simulating a natural level of patience in human-based interactive instruction.

A mechanism for a reading phase is implemented through a finite state machine or equivalent having at least four states which recognizes reading errors at any position in a script and which employs a first set of actions. A related mechanism for an interactive question phase also is implemented through another finite state machine having at least four states, but which recognizes reading errors as well as incorrect answers while invoking a second set of actions.

As part of the linguistically context-sensitive speech recognizer, the probabilistic model of speech is simplified by use of a script for narration, while explicitly modeling disfluencies comprising at least pauses and out-of-script utterances.

In conjunction with the interactive reading and question/answer phases, linguistically-sensitive utterance endpoint detection is provided for judging termination of a spoken utterance to simulate human turn-taking in conversational speech.

A scoring system is provided which is capable of analyzing speech and reading proficiency, i.e., speed and error rate, by weighting the proportion of time during correct reading, the ratio of subject reading speed to nominal native reading speed, and the proportion of "alt" units (a novel model for speech) in recognized word stream.

In connection with a DSP device or an equally-powerful processor, the invention allows for real-time conversation between the system and the user on the subject of a specific lesson. The invention may be used conveniently at a location remote from the system through a telephone network wherein the user accesses the system by selecting a telephone number and references from visual or memorized materials for interaction with the system.

The invention will be better understood by reference to the following detailed description in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a system according to the invention.

Fig. 2 is a functional block diagram of recognition processes employed with the invention.

Fig. 3 is a functional block diagram of processes used in connection with the invention.

Fig. 4A1 is a first portion of a flowchart of a process of pacing a user through a lesson embedded in an apparatus implemented in accordance with the invention.

Fig. 4A2 is a second portion of a flowchart of a process of pacing a user through a lesson embedded in an apparatus implemented in accordance with the invention.

Fig. 4B is a flowchart of a tracking process according to the invention.

Fig. 5 is a state diagram of a sentence-level grammar used in a reading mode according to the invention.

Fig. 6 is a state diagram of a word-level grammar used in accordance with the invention.

Fig. 7 is a state diagram of a sentence-level grammar used in an answering mode according to the invention.

Fig. 8 is a state diagram of an "alt" structure used in the grammars according to the invention.

Fig. 9 is a block diagram of a reading speed calculator.

Fig. 10 is a block diagram of a reading quality calculator.

DESCRIPTION OF SPECIFIC EMBODIMENTS

Referring to Fig. 1, there is shown a system block diagram of an instructional apparatus 10 according to the invention for instructing a user 12 located close to the apparatus 10 or for instructing a user 12' located remotely from the apparatus 10 and communicating via telephone 14. The local user 12 may interact with the system through a microphone 16, receiving instructions and feedback through a loudspeaker or earphones 18 and a visual monitor (CRT) 20. The remote user 12' receives prompts through a published or

printed text 22, as from a newspaper advertisement, or may employ some well-known or memorized text. The remote user's telephone 14 is coupled through a telephone network 24 through a multiplexer 26. The local user's microphone 16 is also coupled to the multiplexer 26. The output of the multiplexer 26 is coupled to a preamplifier 28, through a lowpass filter 30 and then to an analog to digital converter 32, which is part of a digital signal processing (DSP) subsystem 34 in a workstation or timesharing computer 36. Output from the DSP subsystem 34 is provided through a digital to analog converter (DAC) 38 to either or both an amplifier 40 or the telephone network 24, which are respectively coupled to the speaker 18 or the telephone 14. The CRT 20 is typically the visual output device of the workstation 36. A suitable DSP subsystem is the "Sonitech Spirit 30" DSP card, and a suitable workstation is the Sun Microsystems SPARCStation 2 UNIX workstation.

Referring to Fig. 2 in connection with Fig. 1, the basic operation of the underlying system is illustrated. The system is preferably built around a speech recognition system such as the Decipher system of SRI International. The user 12 addresses the microphone (MIC) 14 in response to a stimulus such as a visual or auditory prompt. The continuous speech signal of the microphone 14 is fed through an electronic path to a "front end" signal processing system 42, which is contained primarily in the DSP subsystem 34 and subject to control of the mother workstation 36. The front end signal processing system 42 performs feature extraction, feeding acoustic feature parameters to a model searcher 44 which is built around a hidden Markov Model model set (HMM models) 46. The model searcher 44 performs a "search" on the acoustic features, which are constrained by a finite state grammar to only a limited and manageable set of choices. Hence, significant latitude can be granted the user in quality of pronunciation when compared with the HMM models 46. An application subsystem 48 in the form of a prepared lesson of delimited grammar and vocabulary communicates with the model searcher 44. The application subsystem 48 supplies the finite

state grammar to the model searcher 44 against which a search is performed and the model searcher 44 communicates via backtracing processes embedded in the speech recognition system, such as Decipher, recognition or nonrecognition, as well as backtrace-generated information, to the application subsystem 48, which then interacts with the user 12 according to the invention.

There are two functional modes to a speech processing system used in connection with the invention, a training mode and a recognition mode. The processing is illustrated in reference to Fig. 3. In a training mode, a training script 102 is presented to a plurality of persons in a training population 104, each of which produces a plurality of speech patterns 106 corresponding to the training script 102. The training script 102 and the speech patterns 106 are provided as an indexed set to a hidden Markov model trainer 108 to build general HMM models of target language speech 111. This needs to be done only once for a target language, which typically may employ native speakers and some non-native speakers to generate general HMM models of target language speech. Then an HMM network model compiler 110, using as input the general HMM models and the preselected script 114, builds a network of speech models 113 specifically for the preselected script. The network model compiler output is provided to a hidden Markov model-based speech recognizer 112.

In a recognition mode, a preselected script 114, which is a functional subset of the training script 102 but does not necessarily include the words of the preselected script 102, is presented to a trainee/user 116 or even a device whose pronunciation is to be evaluated. The speech of the trainee/user 116 is presumed to be in the form of a speech pattern 118 corresponding to the preselected script 114. The preselected script 114 and the single speech pattern 118 are provided as an indexed set to the hidden Markov model speech recognizer 112. During each current evaluation period (a phone-length, word-length, phrase-length or even sentence length-period of time), words are recognized by the recognizer 112. From the number of words recognized during the

evaluation period and prior periods, a recognition score set 120 is calculated, passed on to the application subsystem 48 (Fig. 2) serving as a lesson control unit of the type herein described. The score set 120 is a snapshot of the recognition process as embodied in backtrace-generated information. It is passed to the application subsystem 48/lesson control unit which employs a finite state machine embodying the decision apparatus hereinafter explained. The finite state machine, among other functions, filters the raw score set information to identify only good renditions of the scripted lesson. Specifically, it identifies subsets of the score set upon which to judge the quality of lesson performance, including reading speed and reading quality.

Fig. 4A is a flowchart of a process of pacing a user through a lesson embedded in an apparatus implemented in accordance with the invention. It is implemented as a finite state machine (FSM) which is embedded in the application subsystem 48 which controls the interaction of the user 12 and the lesson material.

In operation, reference is directed by the FSM to a script, which may appear on a CRT screen or produced as printed material to be read. Starting with a sentence index of i=1 and a word index j=1 (Step A), a tracking process is executed (Step B). The FSM tests to determine whether the user has finished reading the last sentence in the script (Step C), causing an exit to END if true (Step D). Otherwise the FSM tests to determine whether the user is pausing as detected by the tracker and has read good (recognizable) words from the script since the last tracking operation (Step E). If true, the FSM responds preferably with an aural or visual positive rejoinder, e.g., the response "okay" (Step F), and the FSM recycles to the tracking process (Step B).

If on the other hand, the FSM determines that the user is not pausing after having read good words since the last tracking operation, the FSM prompts the user by stating: "Please read from P(i)." (Step G) The P(i) is the beginning of the identified location in the script of the phrase containing or immediately preceding the untracked words. The

tracking process is thereafter invoked again (Step H), this time at a level of patience wherein the user has effectively one penalty. The FSM then tests for the completion of the last sentence, as before, in this new level (Step I), and ends (Step J) if the script has been completed. Otherwise the FSM tests to determine whether the user is pausing as detected by the tracking operation and has read good (recognizable) words from the script (Step K). If true, the FSM responds with a preferably an aural or visual positive rejoinder, e.g., the response "okay" (Step L), tests for the beginning of a new sentence (Step M) and if yes the FSM recycles to the tracking process (Step B), but if no the FSM recycles to track within the current sentence (Step H).

If words are not being read correctly as indicated by the tracking operation (Step K), the FSM tests to determine whether a new sentence has begun (Step N), in which case the FSM recycles and prompts the user to read from the beginning of the sentence (Step G). If this is not the beginning of a sentence, the FSM states: "No, the sentence is S(i). Please read from P(i)." (Step P). In other words, the user is presented with a model of the sentence and prompted to start at the beginning of the sentence, that is, to try again.

After the prompt, the FSM reinvokes the tracking procedure (Step Q), then tests to see if the last sentence has been spoken (Step R), ending if YES (Step S), otherwise testing to see if the user is pausing after having read good words from the script (Step T). The FSM issues an "ok" if true (Step U), tests for a new sentence (Step V), restarting the tracking (to Step Q) if no, otherwise if a new sentence, resetting to the highest level of patience with tracking (Step B). If the FSM is not tracking good words, it checks to see if a new sentence has started (Step W) and if so, prompts the user tc start reading from the initialize sentence position P(i) (to Step G). If it is not a new sentence, the FSM shows a loss of patience by reciting a phrase such as: "Ok. That was a nice try. Now read from the beginning of the next sentence." (i.e., P(i+1)) (Step Z). The sentence counter index i is then incremented by one sentence (i+1) (Step AA)

and the word counter index j is reset to 1 (Step AB), returning to the initial tracking process (to Step B), where the FSM regains its initial level of patience.

Fig. 4B is a flow diagram of the tracking process (Steps B, H, Q) used by the FSM of Fig. 4A. The tracking process examines one second of input speech (Step AC) using for example a hidden Markov model of speech patterns corresponding to the preselected script. The FSM updates the counters (i & j) to the current position (Step AD) and tests to determine whether the last sentence has been recited (Step AE). If yes, the tracking process is exited (Step AF). If the last sentence is not recognized, the FSM then computes a pause indicator, which is the number of pause phones recognized since the previous word (Step AG), which is in general indicative of the length of a pause. It is then compared with a pause indicator threshold for the current position (i,j) and exercise strictness level (Step AH). If the pause indicator exceeds the threshold, the tracking process is exited (Step AI). If not, the FSM computes a reject indicator (Step AJ). The reject indicator, which is in general indicative of the likelihood that the user is not producing speech corresponding to the preselected script, is computed for instance by summming all reject phones returned by the recognizer since the last word.

The reject indicator is thereafter compared to a reject indicator threshold (Step AK), which is a function of the exercise scoring strictness level or of the current position in the text. If the indicator exceeds the threshold, the procedure is exited (Step AL). If not, a reject density is computed (Step AM).

Reject density is computed by examining a previous number of scripted words (e.g., five) counting the number of reject phones returned by the recognizer, and then dividing the number of reject phones by the sum of the number of reject phone and the number of scripted words (five). That quotient is the reject density. Thus, variations in pause lengths do not impact the reject density.

The reject density is thereafter compared with a reject density threshold (a function of exercise strictness level, text position or both) (Step AN). If the reject density exceeds the threshold, the tracking process is ended (Step AO); otherwise the tracking process is continued (Step AC).

The reject indicator threshold, reject density threshold and pause indicator threshold may be variably adjusted as a function of level of strictness or position in text. The adjusting may be done by the user, by the lesson designer or automatically by the system.

Referring to Fig. 5, there is shown a structure for a sentence-level grammar during the reading phase of the lesson. The sentence level grammar and associated linguistic structures provide the structural sophistication needed to accommodate pauses, hesitation noises and other out-of-script speech phenomenon expected of speech of a student speaker. The grammar consists of "alt" structures 122 separating sentences 126, 128, 130 which have been recognized from the scripted speech patterns. The purpose of the "alt" structure 122 (etc.) is to identify or otherwise account for out-of-script (nonscripted or unscripted) speech or silence (not merely pauses) which is likely to be inserted by the reader into the reading at various points in the reading or answering exercise. An alt structure according to the invention may be used in a hidden Markov model-based speech recognition system to add versatility to a basic speech recognizer enabling it to handle extraneous or unscripted input in an explicit fashion.

Referring to Fig. 6, there is shown the structure of a word-level grammar for a sentence, in either the reading mode or the answering mode. Unlike known word level grammars where a specific key is sought for detection, this grammar explicitly anticipates recitation disfluencies between every word and thus consists of an alt structure 132, 134 between each ordered word 136, 138, each one leading to the next. Whereas words may be returned by the recognizer as atomic units, alt structures are analyzed and returned by the recognizer as strings of reject phones and pause phones which

constitute the alt structures as further detailed herein. This gives the application subsystem 48 (Fig. 2) the ability to render higher-level decisions regarding reading by a user.

Referring to Fig. 7, there is shown the structure of a sentence-level grammar in the answering mode. An initial alt 140 is connected by trajectories to any one of a plurality of answers 142, 144, 146, 148 as alternatives, and each of the answers is connected by trajectories to a final alt 150. This grammar for rejecting unanticipated replies from the user by looping on the initial alt 140, rejecting speech after a valid answer by looping on the final alt 150 or by accepting interjections and pauses during the rendition one of the valid answers.

Fig. 8 illustrates the alt structure 152 common to all alts. The alt structure 152 is a network of hidden Markov states, the parameters of which are trained to account for acoustic features corresponding to out-of-script speech, silence or background noise. It consists of a "pause" model 154 and a "reject" model 156 along alternative forward transition arcs 158, 160, and 162, 164 between an initial node 166 and a terminating node 168. Between the initial node 166 and the terminating node 168 there are also a direct forward transition arc 170 and a direct return transition arc 172. The internal structure of the pause model 154 and the reject model 156 consists of three Markov states and five transition arcs, which is the exact structure used for models of other phones in the Decipher speech recognition system available from SRI International of Menlo Park, California.

The pause model 154 is a phone which is trained on non-speech segments of the training data (typically recorded) and comprises primarily examples of silence or background noise occurring in the training data. The model 156 for the reject phone is a phone which is trained on a wide variety of speech which has been selected randomly or periodically from the training data.

The alt structure 152 with the pause model phone 154 and the reject model phone 156, fully trained, is connected internally by the transition arcs to allow for all of the

following possible events: prolonged silence (multiple loops through the pause phone 154 and the return arc 172); prolonged out-of-script speech (multiple loops through the reject phone 156 and the return arc 172); alternating periods of silence and out-of-script speech; and no pause and no out-of-script speech (bypass on forward transition arc 170).

The initial transition arcs 158 or 162 leading to the pause phone 154 and to the reject phone 156 are in one embodiment of the invention equally weighted with a probability of 0.5 each.

Referring to Fig. 9, there is shown a reading speed calculator 180 according to the invention. It receives from the application subsystem 48 (the finite state machine) a subset (array of data) 182 of the score set 120 identifying the elements of good speech by type (words, pause element, reject element) and position in time, plus certain related timing. Probability information is available but need not be used.

skips leading, trailing silence

Reading speed is extracted by use of a word counter 184, to count the "good" words, and a timer 186, which measures or computes the duration of the phrases containing the filtered (good) words. A reading speed score 190 is determined from a divider 188 which divides the number of "good" words W by the time elapsed T in reciting the accepted phrases containing the "good" words.

The subsystem herein described could be implemented by a circuit or by a computer program invoking the following equations:

Fig. 10 illustrates a mechanism 192 to determining a reading quality score 230. In connection with the system, there is a word count source 194 providing a count value 195 for number of words in the preselected script, a mechanism 196 by which the optimum reading time 197 of the script is reported, a means 198 for counting number of reject phones (199), a means 200 for measuring total time elapsed 201 during reading of all words in the preselected script, and a means 202 for measuring "good" time elapsed 203 during reading of phrases deemed acceptable by said analyzing means.

A divider means 204 is provided for dividing the total time value 201 by the good time value 203 to obtain a first quotient 205, and a weighting means 206 (a multiplier) is providing for weighting the first quotient 205 by a first weighting parameter ("a") to obtain a first score component 208. The sum of three weighting parameters a, b and c is preferably 1.0 by convention to permit an assignment of relative weight of each of three types of quality measurements.

A selector means 210 is provided for selecting a maximum between the optimum reading time 197 and the good time 203 to produce a preferred maximum value 211. This is used in valuing a preference between a fast reading and a reading which is paced according to a preference. In connection with the preference evaluation, a divider means 212 is provided for dividing the preferred maximum value 211 by the optimum reading time 197 to obtain a second quotient 213. The second quotient is weighted by a second weighting parameter (b) by a weighting means 214 (a multiplier) to obtain a second score component 216.

An adder or summing means 218 is provided for summing the number of reject phones 199 and the number of script words 195 to obtain a quality value 219. A divider means 220 is provided for dividing the number of words 195 by the quality value 219 to obtain a third quotient 221. The third quotient is weighted by a weighting means 222 (a multiplier) by third weighting parameter (c) to obtain a third score component 224.

A three-input summing means 226 is provided for summing the first, second and third score components 208, 216 and 224 to produce a score sum 227. The score sum 227 is scaled to a percentage or other scale by a weighting means multiplying by a scale factor 228, such as the value 10 to obtain the reading quality score 230.

The reading quality evaluation subsystem herein described could be implemented by a circuit or by a computer program invoking the following equation:

$$RQS = 10 * (a*T_g/T_t + b*(T_n/[\max (T_n, T_g)]) + c*W/(R_g + W)$$

where:

RQS is the reading quality score on a scale of 1 to 10 (based on the scale factor, herein 10);

a, b, and c are scale factors whose sum equals 1 and in a specific embodiment, a=0.25, b=0.25 and c=0.5;

W is the number of words in the text;

T_g is the "good" time or time spent reading good sentences;

T_t is the total reading time spent reading, excluding initial and final pauses;

T_n is the optimal reading time, i.e., reading time by a good native speaker;

R_g is the number of rejects detected during the "good" renditions of the sentences, i.e., during T_g.

Appendix A is a microfiche appendix of source code listing of a system according to the invention implemented on a computer workstation. The language of the source code is C.

The invention has now been explained with reference to specific embodiments. Other embodiments will be apparent to those of ordinary skill in this art upon reference to the present disclosure. It is therefore not intended that this invention be limited, except as indicated by the appended claims.

WHAT IS CLAIMED IS:

1. In an automatic speech recognition system incorporating a speech recognizer producing word sequence hypotheses and employing a language model for prioritizing a range of word sequence patterns as a constraint on the speech recognizer, a method for tracking a speech pattern and identifying errors in said speech pattern in relation to a preselected script containing alternative texts and interactively prompting a user to recite said preselected script, the method comprising the steps of:

providing to a digital computer a grammar model for a sentence, said grammar model comprising single alt elements disposed between each sequentially-arranged word to form a sentence; and

providing to said digital computer a grammar model for a script by aggregating sentences into strings separated by single alt elements disposed between each sequentially-arranged sentence in a series;

using said speech recognizer trained in a subject language and stored in said digital computer with said grammar models to align speech of a user with strings of words in said script and to identify scripted and nonscripted speech and context-sensitive silence; and

prompting the user in response to said scripted and nonscripted speech and said context-sensitive silence, according to at least three levels of patience, to recite said preselected script with phonetic and semantic accuracy.

2. In the speech recognition system of claim 1, the method further including the step of:

providing a grammar model for alternative texts of sentences, said interactive conversation grammar model comprising a first common alt element disposed before a selection of alternative answers and a second common alt element disposed after said selection of alternative answers, thereby to permit alternative responses having phonetic accuracy and semantic inaccuracy.

3. In the speech recognition system of claim 1, wherein said using step comprises:

recurrently examining a current segment of output of said speech recognizer for scripted words, pause phones and reject phones;

determining reject density for said current segment;

testing said reject density against a reject density threshold; and

denoting speech as out of script if said reject density exceeds said reject density threshold.

4. In the speech recognition system of claim 3, wherein said reject density is determined by dividing number of rejected phones returned by said speech recognizer out of a preselected number of consecutive scripted words by a sum of said rejected phones and said preselected number of words.

5. In the speech recognition system of claim 1, wherein said using step comprises:

recurrently examining a current segment of output of said speech recognizer for scripted words, pause phone and reject phones;

determining a reject indicator for said current segment;

testing said reject indicator against a reject indicator threshold; and

denoting speech as out of script if said reject indicator exceeds said reject indicator threshold.

6. In the speech recognition system of claim 5, wherein said reject indicator determining step comprises summing reject phones returned by said speech recognizer out of a preselected number of consecutive scripted words.

7. In the speech recognition system of claim 1, wherein said using step comprises:

recurrently examining a current segment of output of said speech recognizer for scripted words, pause phones and reject phones;

determining a pause indicator for said current segment;

testing said pause indicator against a pause indicator threshold; and

denoting speech as out of script if said pause indicator exceeds said pause indicator threshold.

8. In the speech recognition system of claim 7, wherein said pause indicator threshold is dependent upon linguistic context and position in text, said pause indicator threshold being smaller at ends of sentences and major clauses than elsewhere among words of sentences.

9. In the speech recognition system of claim 7, wherein said pause indicator determining step comprises summing pause phones returned by said speech recognizer out of a preselected number of consecutive scripted words.

10. In the speech recognition system of claim 2, wherein said alt element is of a structure comprising:

a plurality of transition arcs for events, including prolonged silence;

prolonged out-of-script speech;

alternating periods of silence and out-of-script speech; and

no pause and no out-of-script speech.

11. A system for tracking speech of a user with spoken inputs to the system and spoken and graphic outputs using an automatic speech recognition subsystem incorporating a speech recognizer producing word sequence hypotheses and employing a language model for prioritizing a range of word sequence patterns as a constraint on the speech recognizer, the system comprising:

means for presenting information to the user about a subject and inviting a reading of a preselected script of allowable utterances;

means for sensing an acoustic signature indicative of a speech-containing signal from a time-invariant frame of acoustic information;

means for analyzing said frame of acoustic information to determine a set of possible utterances corresponding to an accumulation of acoustic information frames;

means coupled to said analyzing means for assessing completeness of an utterance to determine accuracy of reading; and

means coupled to said comparing means for producing a response encouraging correct reading of the preselected script.

12. The system according to claim 11 wherein the tracking system is for instruction in a language foreign to the user and wherein said producing means includes means for generating an audible response as an example of native pronunciation and rendition.

13. In the system according to claim 11, further including means for measuring reading speed comprising:

means for counting number of words read;

means for measuring time elapsed during reading scripted words; and

means for dividing said number of words counted by said measured time elapsed.

14. In the system according to claim 11, further including means (192) for measuring reading quality to obtain a reading quality score (230) comprising:

means (194) providing a count for number of words (195) in the preselected script;

means (196) providing a time duration establishing an optimum reading time (197);

means (198) for counting number of reject phones (199);

means (200) for measuring total time elapsed (201) during reading of all words in the preselected script;

means (202) for measuring good time elapsed (203) during reading of phrases deemed acceptable by said analyzing means;

means (204) for dividing said total time (201) by said good time (203) to obtain a first quotient (205);

means (206) for weighting said first quotient (205) by a first weighting parameter (a) to obtain a first score component (208);

means (210) for selecting a maximum between said optimum reading time (197) and said good time (203) to produce a preferred maximum value (211);

means (212) for dividing said preferred maximum value (211) by said optimum reading time (197) to obtain a second quotient (213);

means (214) for weighting said second quotient (213) by a second weighting parameter (b) to obtain a second score component (216);

means (218) for summing said number of reject phones (199) and said number of words (195) to obtain a quality value (219);

means (220) for dividing said number of words (195) by said quality value (219) to obtain a third quotient (221);

means (222) for weighting said third quotient (221) by a third weighting parameter (c) to obtain a third score component (224);

means (226) for summing said first score component (208), said second score component (216) and said third score component (224) to produce a score sum (227); and

means for weighting said score sum (227) by a scale factor (228) to obtain said reading quality score (230).

15. A system for tracking speech and interacting with a user with spoken inputs to the system and spoken and graphic outputs using an automatic speech recognition

subsystem incorporating a speech recognizer producing word sequence hypotheses and employing a language model for prioritizing a range of word sequence patterns as a constraint on the speech recognizer, the system comprising:

means for presenting information to the user about a subject and inviting a rejoinder from a preselected set of allowable utterances to evoke a spoken response;

means for sensing an acoustic signature indicative of a speech-containing signal from a time-invariant frame of acoustic information;

means for analyzing said frame of acoustic information to determine a set of possible utterances corresponding to an accumulation of acoustic information frames;

means coupled to said analyzing means for assessing completeness of an utterance from said set of utterances;

means coupled to said assessing means for selecting a best hypothesis for an utterance from said set of possible utterances upon indication of the end of an utterance;

means coupled to said selecting means for comparing said best hypothesis with the preselected set of allowable utterances to determine the rejoinder selected; and

means coupled to said comparing means for producing a response corresponding to the rejoinder selected.

16. The system according to claim 15 wherein the interacting system is for instruction in a language foreign to the user and wherein said producing means includes means for generating an audible response as an example of native pronunciation and rendition.

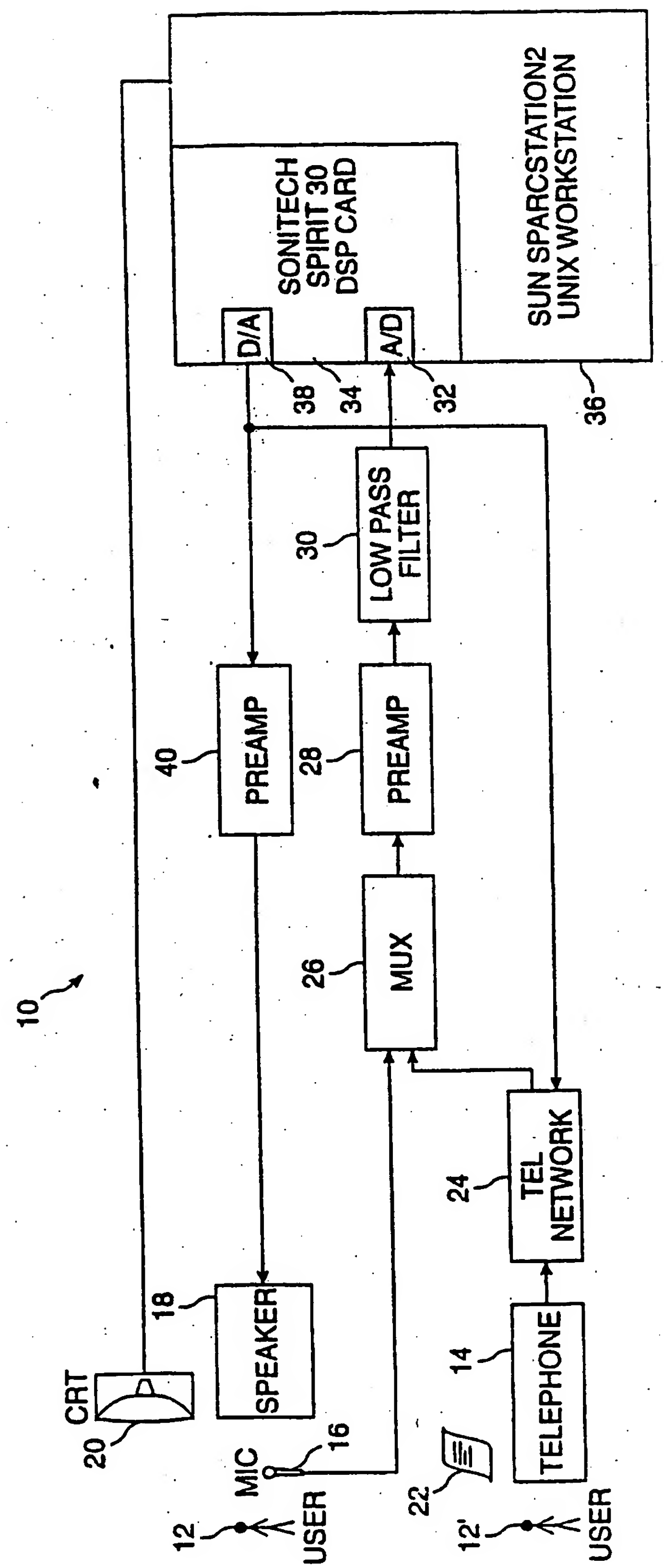

SUN SPARCSTATION2
UNIX WORKSTATION
SONITECH
SPIRIT 30
DSP CARD
D/A
A/D
38
34
32
36
LOW PASS
FILTER
30
PREAMP
40
PREAMP
28
MUX
26
10
TEL
NETWORK
24
SPEAKER
18
CRT
20
MIC
16
TELEPHONE
14
22
12
USER
12'
USER

FIG. 1

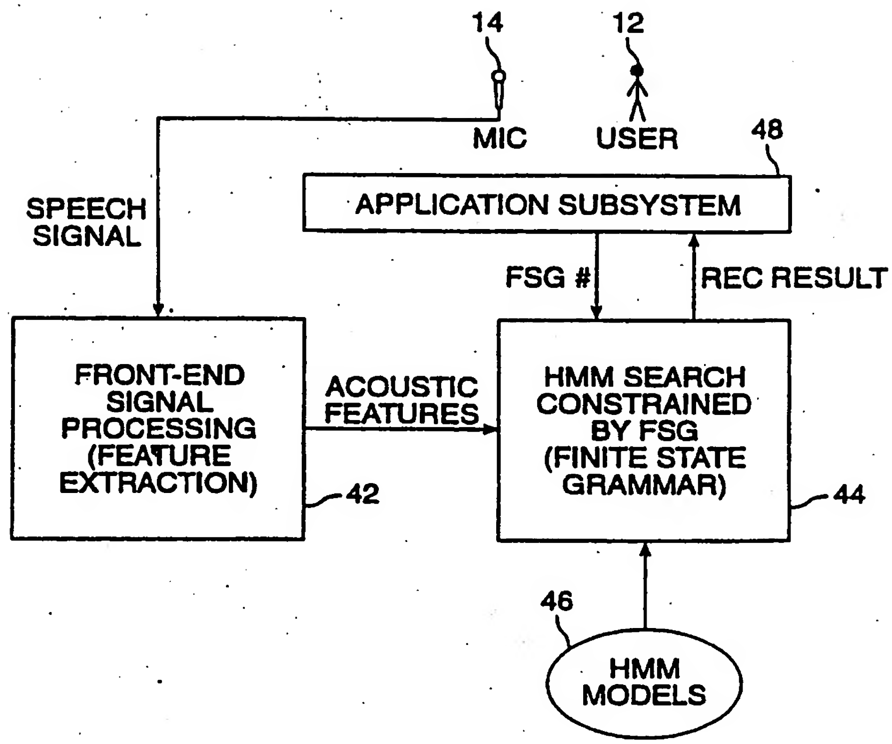

14
12
MIC
USER
48
APPLICATION SUBSYSTEM
SPEECH SIGNAL
FSG #
REC RESULT
FRONT-END SIGNAL PROCESSING (FEATURE EXTRACTION)
ACOUSTIC FEATURES
HMM SEARCH CONSTRAINED BY FSG (FINITE STATE GRAMMAR)
42
44
46
HMM MODELS

FIG. 2

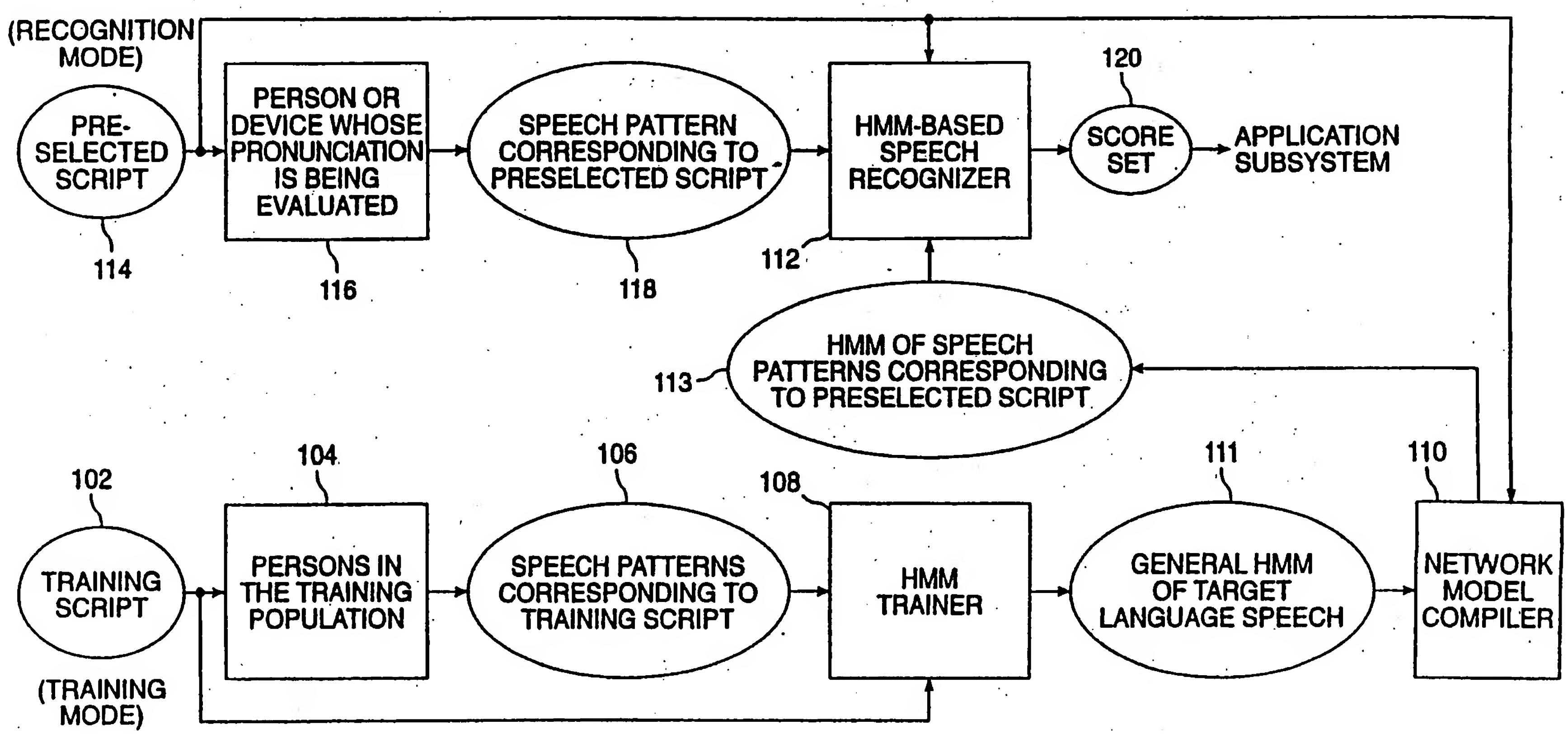
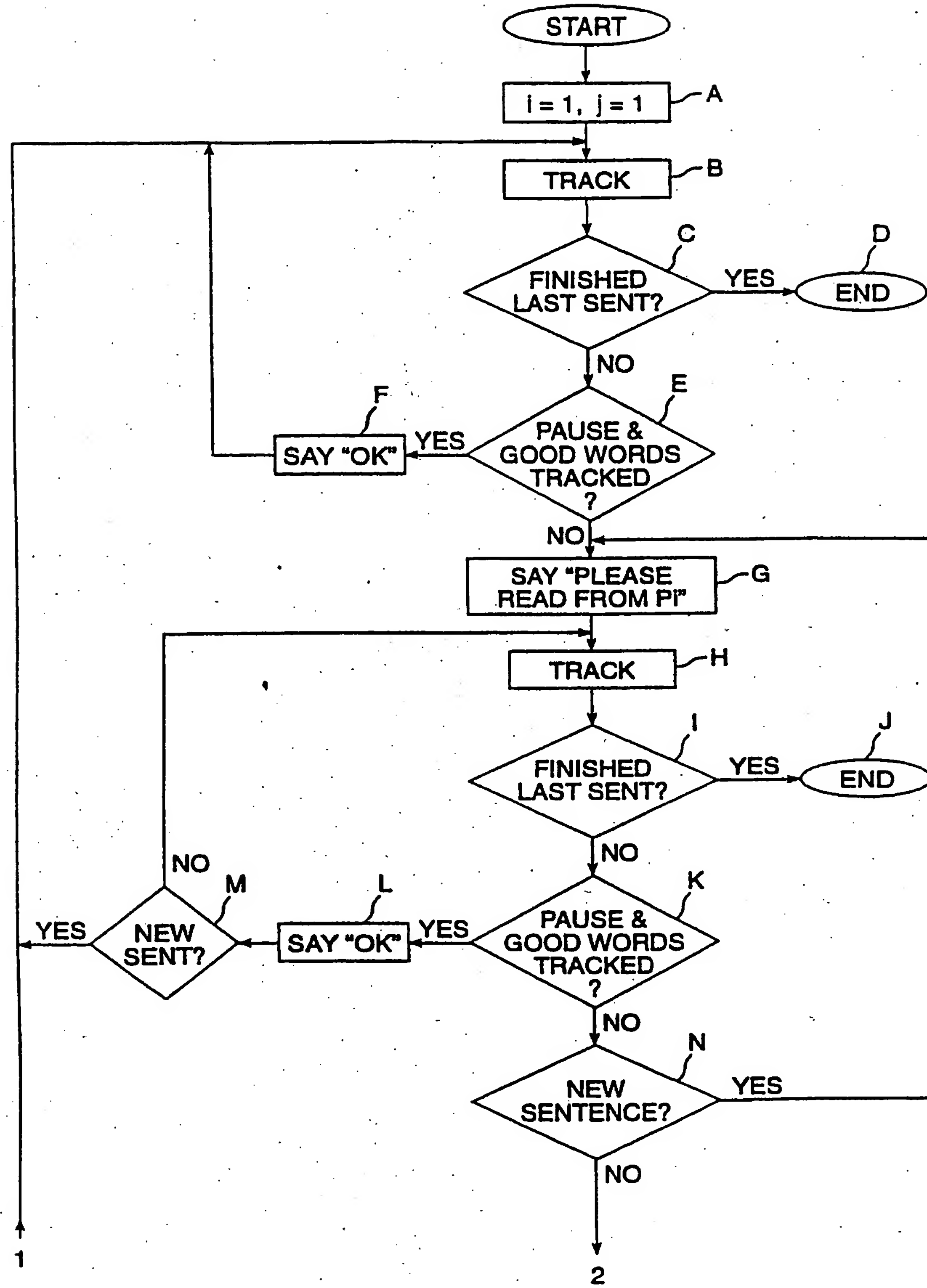

(RECOGNITION MODE)
PRE-SELECTED SCRIPT
114
PERSON OR DEVICE WHOSE PRONUNCIATION IS BEING EVALUATED
116
SPEECH PATTERN CORRESPONDING TO PRESELECTED SCRIPT
118
HMM-BASED SPEECH RECOGNIZER
112
SCORE SET
120
APPLICATION SUBSYSTEM
HMM OF SPEECH PATTERNS CORRESPONDING TO PRESELECTED SCRIPT
113
TRAINING SCRIPT
102
(TRAINING MODE)
PERSONS IN THE TRAINING POPULATION
104
SPEECH PATTERNS CORRESPONDING TO TRAINING SCRIPT
106
HMM TRAINER
108
GENERAL HMM OF TARGET LANGUAGE SPEECH
111
NETWORK MODEL COMPILER
110

FIG. 3



START
i = 1, j = 1
A
TRACK
B
C
D
FINISHED LAST SENT?
YES
END
NO
E
F
PAUSE & GOOD WORDS TRACKED ?
YES
SAY "OK"
NO
SAY "PLEASE READ FROM Pi"
G
TRACK
H
I
J
FINISHED LAST SENT?
YES
END
NO
NO
M
L
K
YES
NEW SENT?
SAY "OK"
YES
PAUSE & GOOD WORDS TRACKED ?
NO
N
NEW SENTENCE?
YES
NO
1
2
3

FIG. 4A1

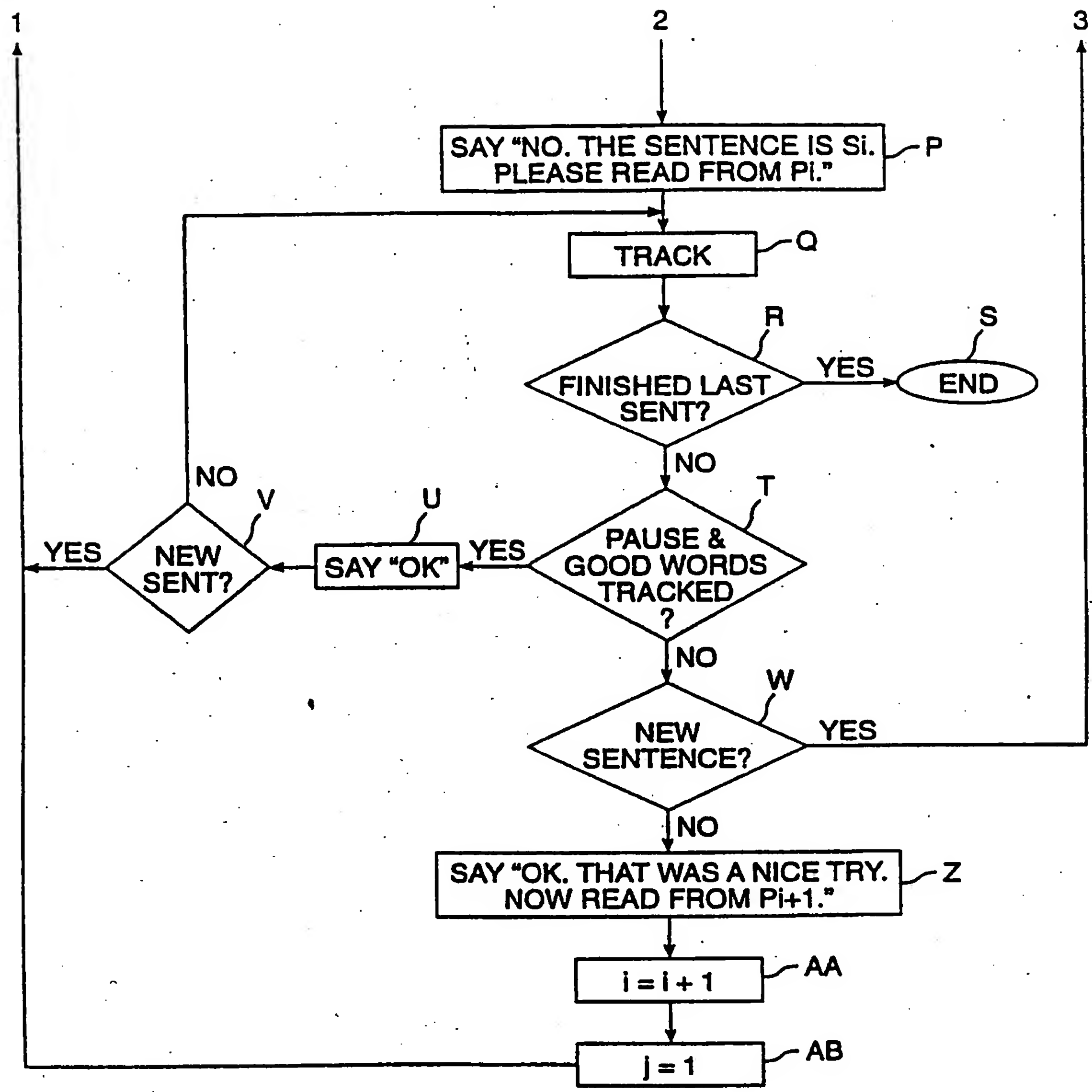

1
2
3
SAY "NO. THE SENTENCE IS Si. PLEASE READ FROM Pi."
P
TRACK
Q
R
FINISHED LAST SENT?
YES
S
END
NO
T
PAUSE & GOOD WORDS TRACKED ?
YES
SAY "OK"
U
NEW SENT?
V
NO
YES
NO
W
NEW SENTENCE?
YES
NO
SAY "OK. THAT WAS A NICE TRY. NOW READ FROM Pi+1."
Z
i = i + 1
AA
j = 1
AB

FIG. 4A2

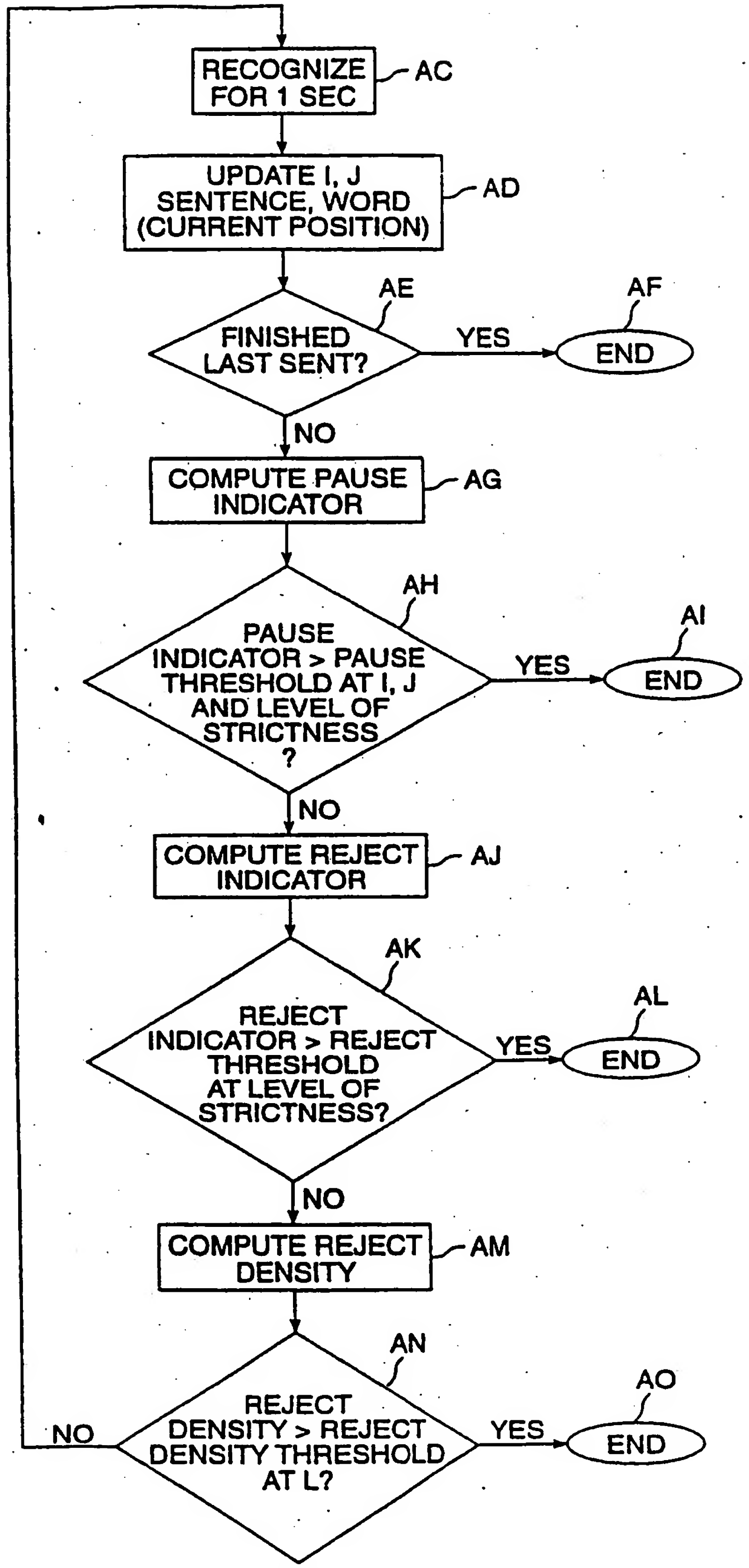

RECOGNIZE FOR 1 SEC
AC
UPDATE I, J SENTENCE, WORD (CURRENT POSITION)
AD
AE
FINISHED LAST SENT?
YES
AF
END
NO
COMPUTE PAUSE INDICATOR
AG
AH
PAUSE INDICATOR > PAUSE THRESHOLD AT I, J AND LEVEL OF STRICTNESS ?
YES
AI
END
NO
COMPUTE REJECT INDICATOR
AJ
AK
REJECT INDICATOR > REJECT THRESHOLD AT LEVEL OF STRICTNESS?
YES
AL
END
NO
COMPUTE REJECT DENSITY
AM
AN
REJECT DENSITY > REJECT DENSITY THRESHOLD AT L?
NO
YES
AO
END

FIG. 4B

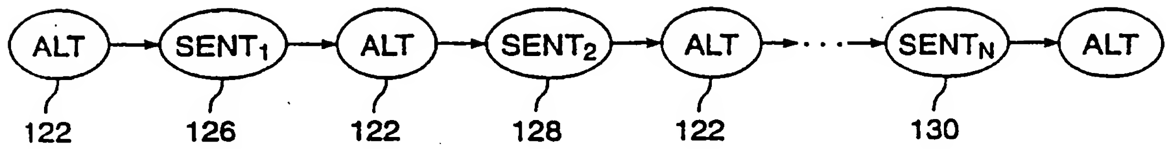


FIG. 5

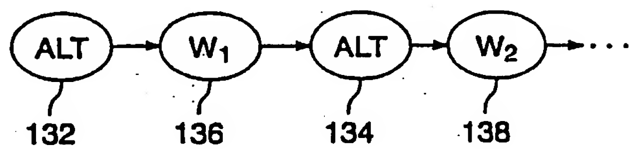


FIG. 6

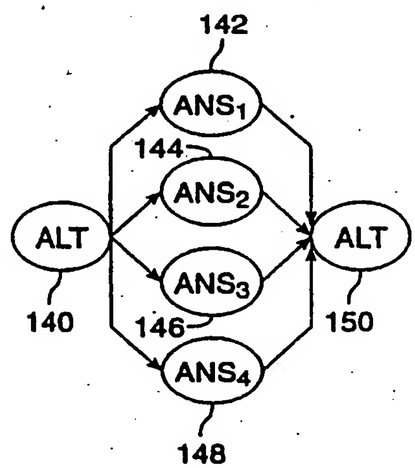


FIG. 7

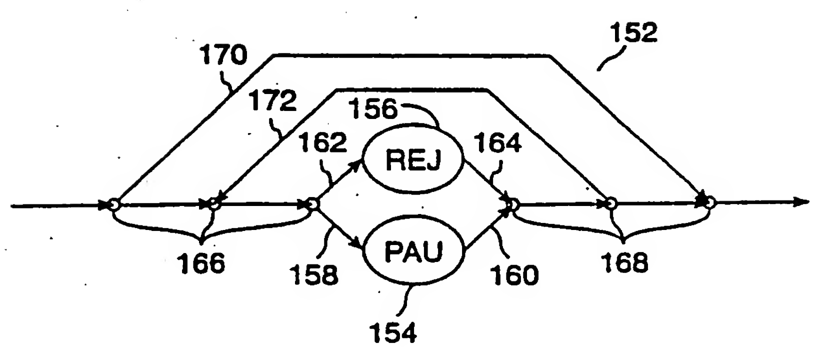


FIG. 8

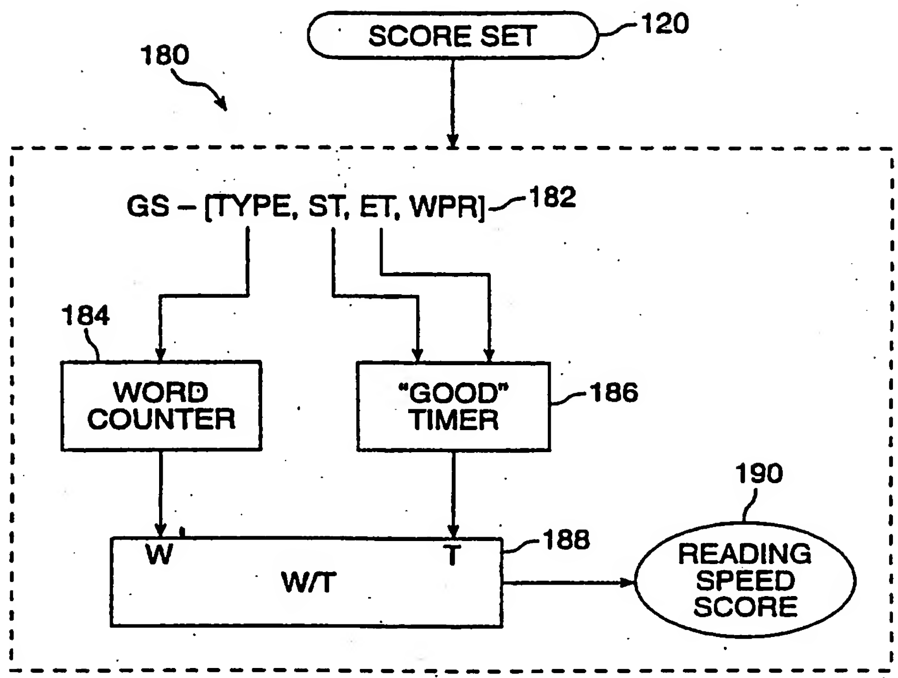

SCORE SET
120
180
GS – [TYPE, ST, ET, WPR]
182
184
WORD COUNTER
"GOOD" TIMER
186
190
W
T
W/T
188
READING SPEED SCORE

FIG. 9

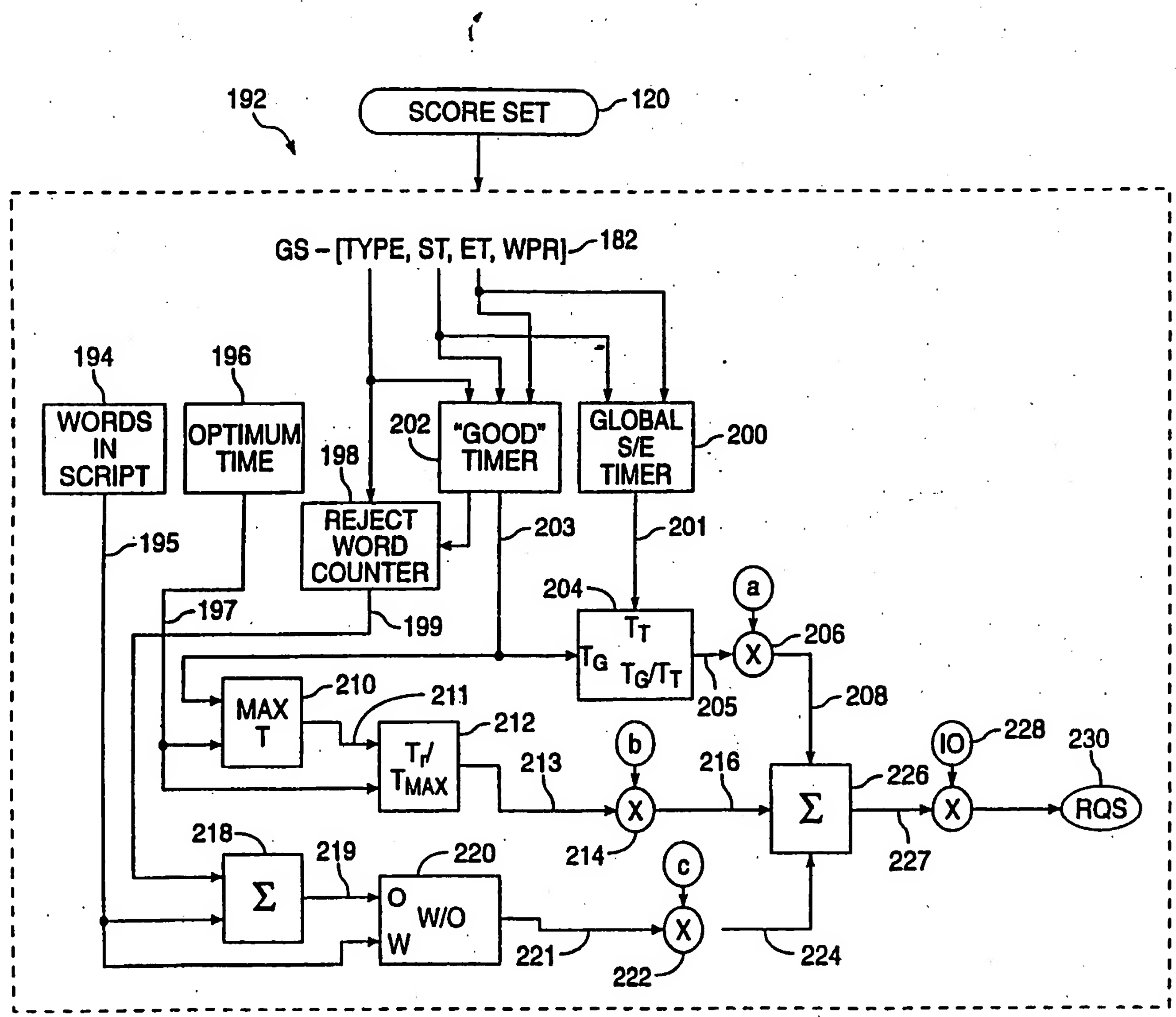

192
SCORE SET
120
GS – [TYPE, ST, ET, WPR]
182
194
WORDS IN SCRIPT
196
OPTIMUM TIME
198
REJECT WORD COUNTER
202
"GOOD" TIMER
GLOBAL S/E TIMER
200
195
197
199
203
201
204
T_T
T_G
T_G/T_T
205
a
X
206
208
MAX T
210
211
T_r/T_{MAX}
212
213
b
214
216
Σ
226
IO
228
230
RQS
227
218
219
220
O
W/O
W
221
c
222
224

FIG. 10

IPC(5) :G10L 9/00
US CL : 395/2.65, 2.64, 2.66, 2.6, 2.54
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 395/2.65, 2.64, 2.66, 2.6, 2.54

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

none

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

none

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US, A, 4,852,180 (Levinson) 25 JULY 1989, abstract.	1-16
Y	US, A, 4,972,485 (Dautrich et al) 20 November 1990, Figure 3 and abstract.	1-16
Y	US, A, 5,033,087 (Bahl et al) 16 July 1991, abstract.	1-16
Y	US, A, 5,031,217 (Nishimura) 09 July 1991, Figures 1 and 7.	1-16

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be part of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
09 MAY 1994	13 JUN 1994
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer ALLEN MACDONALD for Telephone No. (703) 305-9646